**UNIVERSIDAD DE SAN CARLOS DE GUATEMALA**

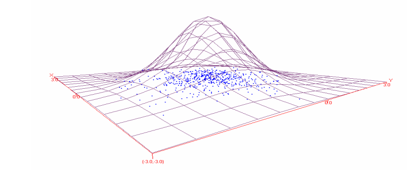
**FACULTAD DE AGRONOMIA**

**CENTRO DE TELEMATICA**

**SUBAREA DE METODOS DE CUANTIFICACION E INVESTIGACIÓN**

**PROGRAMA PERMANENTE DE CAPACITACION**

**EN ESTADÍSTICA APLICADA**



# MODELACIÓN DE REGRESIÓN

M.C. Victor Manuel Alvarez Cajas[1]

Ing. Agr. Byron Humberto González Ramírez[2]

[1] Ingeniero Agrónomo en Sistemas de Producción Agrícola (FAUSAC, Guatemala) Maestro en Estadística Aplicada por el Colegio de Postgraduados, México. Coordinador de la Subárea de Métodos de Cuantificación e Investigación , Facultad de Agronomía USAC. Guatemala.

[2] Ingeniero Agrónomo en Sistemas de Producción Agrícola (FAUSAC, Guatemala). Maestrando en Nuevas Tecnologías de la Información y Comunicación- UNED España- Director del Centro de Telemática, Facultad de Agronomía, USAC. Guatemala.

Centro de Telemática
Facultad de Agronomía
USAC



# MODELACION DE REGRESIÓN

## 1. INTRODUCCIÓN

Con la disponibilidad amplia del uso de las computadoras y el acceso al software estadístico la modelación estadística ha logrado un desarrollo considerable. En esto se incluye el manejo amplio del volumen de datos, lo que ha permitido que las tareas de cálculo manual estén relegadas a la prehistoria estadística.

En la actualidad se puede pensar en realizar un análisis iterativo, es decir, un "diálogo" entre el investigador y el asistente electrónico estadístico. Consecuencia de esta relación será una mejor valoración de las técnicas estadísticas.

Por tal razón es conveniente hacer más énfasis en los conceptos clave y aspectos metodológicos, que en las tareas mecánicas de cálculo. Es claro que una vez obtenidos los resultados que proporcione el computador, el usuario investigador debe saber interpretar y traducir al contexto del problema los mismos que aparecen impresos.

En estas notas se pretende dar lineamientos que permitan la modelación de los fenómenos que sean descritos por los modelos de regresión lineal simple.

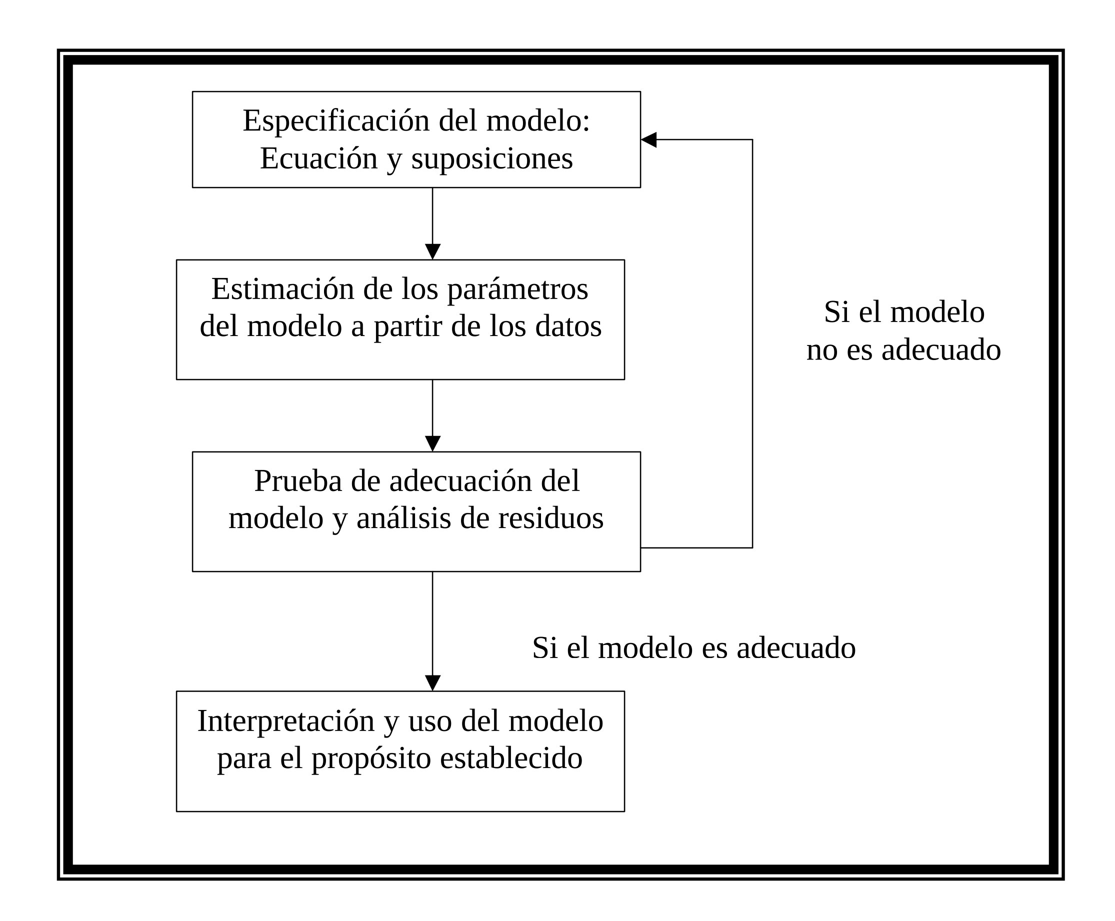
## 2. MODELACIÓN ESTADISTICA

El término modelo fue empleado originalmente por Niels Bohr en la descripción del átomo publicada en 1913, y a partir de la fecha, el uso y la generalización del concepto ha sido considerada como uno de los máximos logros del desarrollo humano.

De manera amplia un modelo es la explicación teórica de un fenómeno bajo estudio. Esta explicación se bosqueja en forma verbal y luego se formaliza mediante una o más ecuaciones.

En la fase de postulación del modelo, desde luego, el criterio práctico y los conocimientos sobre el fenómeno de estudio son fundamentales, una sugerencia de la secuencia de las fases de modelación se presenta en el Cuadro 1.

**Cuadro 1. Fases en la construcción de un modelo estadístico**




## 3. EL MODELO DE REGRESIÓN LINEAL

Alrededor de 1860, Sir Francis Galton[3] (1822-1911), estudiando la relación existente entre las alturas de padres e hijos, llegó al primer modelo conocido ahora como "regresión". De hecho, este nombre se debe al propio Galton, quien describía la relación existente entre dichas alturas como una "regresión a la mediocridad" o llamado por otros autores como la "Ley de la Regresión Filial", indicando con esto

[3] Antropólogo Británico que además de sus invaluables aportes a la Herencia y a la Estadística, fue quien diseñó el sistema de identificación de los individuos humanos con base a la irrepetibilidad de la huellas digitales.

que los padres altos tenían hijos ya no tan altos y los padres bajos tenían hijos ya no tan bajos, existiendo una tendencia a regresar hacia los valores medios.

La Regresión es una técnica estadística que se utiliza para investigar y modelar las relaciones entre variables, con el propósito de usar la información que proporciona una de ellas para tratar de conocer en forma aproximada el comportamiento de la otra. El beneficio que se deriva de llevar a cabo un procedimiento como el expuesto anteriormente es de diversos órdenes, por ejemplo, puede ser mas económico o práctico observar (medir) una característica que otra en tal sentido, seria conveniente poder "predecir" valores de la variable que presenta problemas con base en la observación de la otra. Por ejemplo, en un estudio se desea conocer el grado de conocimiento político de las personas, el cual puede ser difícil de medir, sin embargo, otra variable muy relacionada puede medirse tal como, años de educación, edad, o ingreso.

En ocasiones el uso del modelo de regresión es útil por la estimación de sus parámetros que en algunas ciencias es de interés particular, como por ejemplo, el coeficiente de elasticidad en las ciencias económicas.

## 4. ALGUNAS DEFINICIONES Y NOTACION

Cuando se tiene un grupo de variables para su análisis, se tiene que identificar a las variables que intervienen, es decir, a la variable que se usará para estimar a otra variable se llamará **predictora, regresora, explicatoria o independiente**, la cual es denotada por la letra **X**. Por lo tanto, la variable que será estimada es llamada **variable respuesta o dependiente**, denotada como **Y**. De otro modo, para un economista interesado en predecir el precio de una vivienda en base al número de metros cuadrados construidos, para este caso, la variable X será el número de metros cuadrados construidos y la variable Y será el precio de la misma.

Si la predicción de la variable Y se hace considerando solo una X, la **regresión es simple** , si fueran mas de dos X ($X_1$, $X_2$, . . .), sería el caso de la **regresión múltiple**.

## 5. REGRESION Y CAUSALIDAD

El estudio de las relaciones entre variables ocupa un lugar preponderante en la ciencia moderna, en buena parte porque después de la revolución científica la idea de “causa eficiente”, en el sentido de algo que al manifestarse produce un efecto reconocible, es una idea central del llamado “método científico”. De acuerdo con esto, no es extraño que la existencia de una relación funcional entre dos variables a menudo se interprete de manera mecánica como una relación de causa efecto. Es decir, que no se puede inferir este tipo de relación del mero análisis de regresión.
El análisis de regresión puede ayudar a confirmar una relación causa-efecto, pero no puede ser la base única de tal reclamo.
Un ejemplo muy frecuentemente citado en los tratados clásicos, es aquel en donde se encontró una relación directa muy estrecha entre el número de nacimientos y el número de cigüeñas presentes en una ciudad de Europa. Obviamente esta relación no puede ser considerada como causal, y lo mas seguro es que sólo se deba al azar.

Finalmente es importante apuntar que el análisis de regresión es un instrumento que proporciona una aproximación en la descripción de una relación causa-efecto, la cual está relacionada a la solución del problema, es decir, la ecuación de regresión por si misma no puede ser el objetivo primario del estudio. El contexto del problema real y su adecuada comprensión y planteamiento es lo más importante.
Los datos usados en un análisis de regresión deben ser representativos de la población bajo estudio. Sin datos representativos, el modelo de regresión, y en consecuencia las conclusiones obtenidas de éste, podrían tener graves errores.

## 6. USOS DE LA REGRESIÓN

Usualmente, una inquietud, un problema de investigación o una necesidad específica dan origen a la realización de un estudio en cuyo contexto se debe ajustar y evaluar un modelo de regresión. De hecho, el análisis de regresión puede ser considerado como la técnica estadística mas ampliamente utilizada.

Las aplicaciones del análisis de regresión son numerosas y ocurren en casi todos los campos, en donde se incluye, las ciencias sociales, ciencias económicas, ciencias físicas, ciencias biológicas, ciencias de la salud, entre otras. Algunos ejemplos son presentados a continuación.

**Ejemplo 1.**

En Educación, se puede tener interés en conocer la relación entre el número de horas que un estudiante duerme y su promedio de calificaciones.

**Ejemplo 2.**

En Finanzas, se desea construir un modelo de regresión que pueda predecir la cantidad que se pueda recaudar mensualmente en función de la cantidad de dinero que se invierte en publicidad.

**Ejemplo 3.**

En Medicina, se desea realizar un estudio para cuantificar la relación entre la pérdida diaria de lípidos con la pérdida de energía en las heces de niños con fibrosis quística.

**Ejemplo 4.**

En estudios de Contaminación, se puede tener interés en mostrar la relación que pueda existir entre la tasa de mortalidad humana y la contaminación por sulfatos.

**Ejemplo 5.**

En Ciencias Marinas se desea conocer la forma de la relación que existe entre la longitud de una especie y su edad.

**Ejemplo 6.**

En Vulcanología, se puede tener interés en predecir el intervalo de la siguiente erupción de un volcán a partir de la duración de la última erupción.

**Ejemplo 7.**

En Economía, se tiene interés en estudiar la relación entre el valor de la tierra y la distancia a la carretera pavimentada.

**Ejemplo 8.**

En agronomía, se realizan estudios para describir la forma de la relación entre la producción de un cultivo con la cantidad de fertilizante aplicado.

**Ejemplo 9.**

En estudios económicos diversos, la regresión puede tener utilidad para poder predecir el ingreso mensual de un profesor universitario en función de los artículos publicados.

**Ejemplo 10.**

En Medicina Veterinaria, se hacen estudios de la relación entre la edad del animal con la presión arterial observando por separado a los machos y las hembras de esa especie.

## EJERCICIO 1

1. Cite 10 ejemplos en su campo principal de actividad en los cuales la variable dependiente (Y) pueda ser predicha por otra variable (X).
2. En los siguientes ejemplos, establezca e identifique cuales son las variables dependientes (Y), y cuales son las variables regresoras (X).
   a. El gerente de una cadena de supermercados desea investigar la relación entre el número de empleados y las ventas semanales.
   b. En un experimento para estudiar el efecto de la exposición a bajas temperaturas sobre el bacilo de la fiebre tifoidea se expusieron cultivos del bacilo durante diferentes periodos de tiempo a una temperatura de $-5^{o}$ C. Se anotaron los valores de número de semanas de exposición y porcentaje de bacilos sobrevivientes.
   c. En un hospital pediátrico se realizó un estudio para determinar la relación entre el peso y la altura de 120 niños de la misma edad.
   d. En un estudio se trató de determinar el tiempo durante el cual la madera seca absorbe agua. Para este fin se eligió una pieza de madera y se le mantuvo sumergida en agua durante 280 días, pesando la madera a diferentes tiempos para determinar la cantidad de agua absorbida.
   e. En un experimento se estudió el efecto de la exposición a la luz de una enzima. Las variables medidas fueron horas de exposición y la

actividad enzimática relativa (%), con respecto a cero horas de exposición.

f. Un genetista está estudiando el efecto del número de horas de exposición en la incidencia de mutaciones en la población de la mosca de las frutas.

g. Una empresa está considerando introducir un nuevo producto a varias ciudades. Desea establecer si varia el volumen de ventas del producto al variar el número de comerciales de televisión.

h. Se realizó un estudio para observar el efecto del precio de la naranja sobre la cantidad vendida.

i. Para determinar que características influyen mas sobre el precio de venta de una casa se tomaron los siguientes datos: número de habitaciones, área total construida y tiempo de estar construida.

j. El gerente de una empresa desea investigar la relación entre el número de días no autorizados que los empleados no asisten a su trabajo y la distancia entre su hogar a la empresa.

## 7. FORMULACION ADECUADA DEL MODELO.

Para tener una primera impresión de la relación entre las variables se sugiere elaborar un **gráfico o diagrama de dispersión.** El gráfico de dispersión consiste en un plano formado por dos ejes, en los cuales los valores de cada pareja de datos son localizados. La elaboración de tal gráfico es sencilla, enseguida se describen cuatro puntos necesarios:

i. **Identificación.** El diagrama de puntos debe tener un título informativo. Los ejes también deben estar identificados con el nombre de las variables y las unidades en que fueron medidas. Es costumbre como en álgebra, colocar en el eje horizontal a la variable que se considera independiente o regresora (X), y la variable respuesta (Y) en el eje vertical.

ii. **Escala.** Las unidades en ambos ejes deben ser cuidadosamente seleccionadas. La escala para los dos ejes puede no ser la misma. Otro detalle importante es, que ambos ejes no necesariamente se interceptaran en el origen (cero), cuando esto suceda, es necesario cortar el o los ejes con dos líneas paralelas inclinadas, así: \\ para el eje X, y para el eje vertical: //

iii. **Los puntos.** Los puntos en el gráfico deben estar solos, no se les debe unir con línea alguna. El propósito es que proporcionen una impresión visual de la forma de la relación entre ambas variables.

iv. **El cuadrante.** Generalmente el cuadrante que se observa es el que muestra los valores positivos de los ejes X e Y. Pero también, para otros casos los valores negativos pueden observarse, por ejemplo, cuando ocurren pérdidas o decrecimientos.

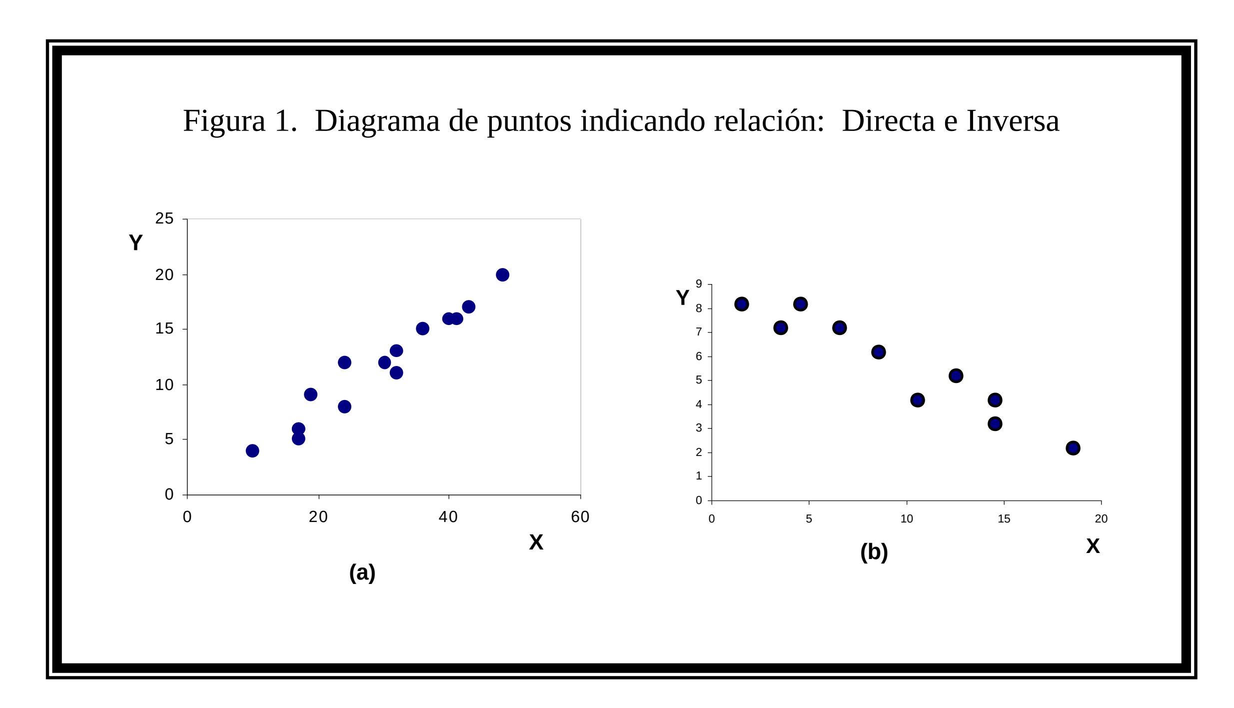
## UTILIDAD DEL DIAGRAMA DE PUNTOS.

El diagrama de puntos permite observar si existe relación lineal o no, en ese sentido, pueden suceder los siguientes casos:

a. **Relación directa.** Se dice que existe una relación directa o positiva entre X e Y, cuando se observa que a valores grandes de X le corresponden valores también grandes de Y. Es decir, por ejemplo, mientras mas grande sea el área de construcción de una casa, su valor de venta también aumentará. Figura 1(a).

b. **Relación inversa.** Otra forma de la relación se puede dar cuando a valores grandes de X, le corresponden valores pequeños de Y, o viceversa. Si ésto sucede, se dice que la relación es inversa o negativa. Por ejemplo, para ciertos productos, cuando el precio disminuye la demanda aumenta y viceversa. Figura 1(b).

Figura 1. Diagrama de puntos indicando relación: Directa e Inversa




c. **No hay relación.** Si al observar el gráfico ninguna de las tendencias anteriores se distingue, visualmente se puede llegar a la conclusión que no hay relación entre X e Y. Para este caso, puede darse que si se aumenta o disminuye X, los valores de la variable Y se mantienen constantes, este comportamiento puede interpretarse, según el contexto del problema, para un economista será el caso clásico de la demanda inelástica , para un matemático será, la cantidad demandada es constante en función del precio. Y para un estadístico esto significaría que las dos variables no están relacionadas. Figura 2(a). Note que el diagrama de puntos está casi paralelo al eje de las X´s. Otra posibilidad sería, que entre los puntos del plano no se distinga comportamiento alguno, este caso entonces, es el ejemplo de la no relación total entre las variables. Figura 2(b).

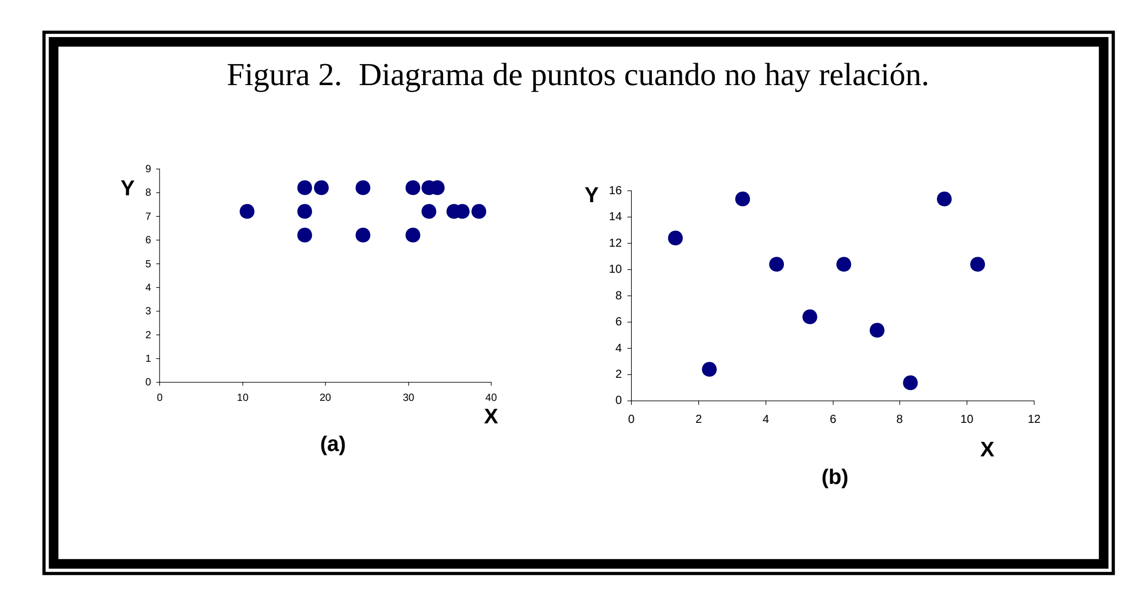



Figura 2. Diagrama de puntos cuando no hay relación.

Otro beneficio importante del diagrama de puntos consiste en permitir identificar si la relación entre las variables es **lineal** o no. Para que sea lineal la relación las Figuras 1(a) y 1(b) son típicas, es decir, que los puntos pueden encerrarse entre dos líneas paralelas. En caso contrario, se deberá buscar otro modelo más adecuado.

Ejemplo.

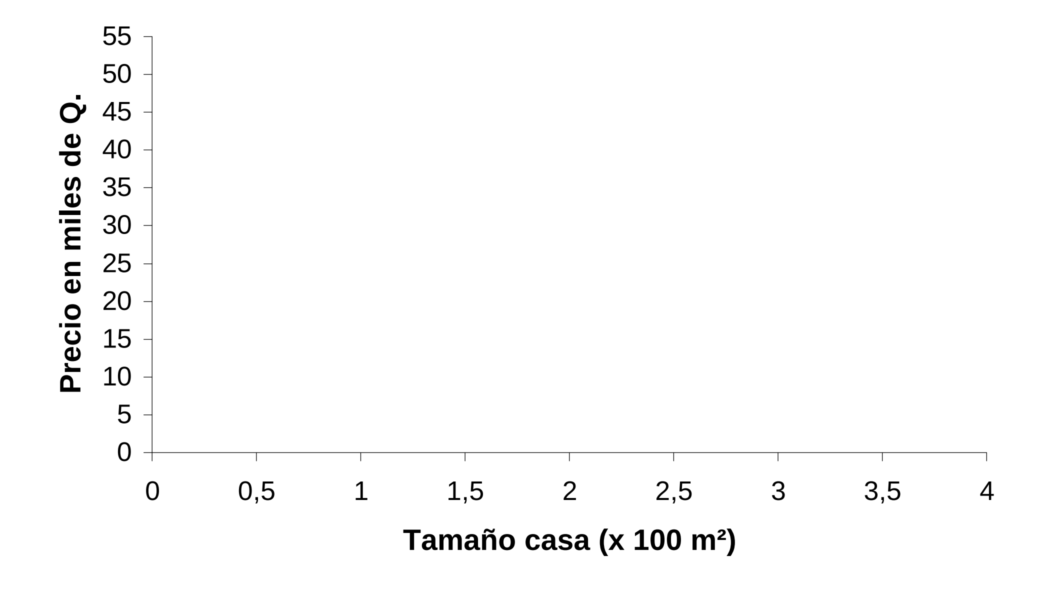
Asuma que una muestra aleatoria de 20 casas tienen los precios y tamaños que aparecen en la Tabla 1.

**Tabla 1. Precios y tamaños de 20 casas.**

| No. Casa | Tamaño (x 100 m²) | Precio en miles de Q. | No. Casa | Tamaño (x 100 m²) | Precio en miles de Q. |
|---|---|---|---|---|---|
| **1** | 1.8 | 32 | **11** | 2.3 | 44 |
| **2** | 1.0 | 24 | **12** | 0.9 | 19 |
| **3** | 1.7 | 27 | **13** | 1.2 | 25 |
| **4** | 2.8 | 47 | **14** | 3.4 | 50 |
| **5** | 2.2 | 35 | **15** | 1.7 | 30 |
| **6** | 0.8 | 17 | **16** | 2.5 | 43 |
| **7** | 3.6 | 52 | **17** | 1.4 | 27 |
| **8** | 1.1 | 20 | **18** | 3.3 | 50 |
| **9** | 2.0 | 38 | **19** | 2.2 | 37 |
| **10** | 2.6 | 45 | **20** | 1.5 | 28 |

Tomando los valores de la Tabla 1, elabore el correspondiente diagrama de puntos en el cuadrante que aparece a continuación:

**Figura 3. Diagrama de puntos para tamaño y precio de las 20 casas**




Del diagrama de dispersión se observa que existe una relación lineal directa entre el tamaño de la casa y el precio de venta, es decir, cuando una casa

tiene mayor cantidad de metros cuadrados de construcción su precio será mayor.

**EJERCICIO 2.**

2.1 A continuación se presentan cuatro conjuntos de datos. Los datos fueron artificialmente construidos y el tamaño de cada serie es pequeño para darle facilidad al ejercicio. Para cada conjunto de datos presentados se pide:

a. Elaborar el diagrama de dispersión.

b. Identificar si la relación es inversa, directa o no existe.

c. Indicar si la relación es lineal o no.

**Conjunto 1.**

El gerente de una empresa desea investigar la relación entre el número de días no trabajados sin autorización por año y la distancia (km) de los empleados. Una muestra de 10 empleados fue seleccionada, y los datos colectados fueron los siguientes:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Distancia al trabajo (km) | 1 | 3 | 4 | 6 | 8 | 10 | 12 | 14 | 14 | 18 |
| Número de ausencias | 8 | 7 | 8 | 7 | 6 | 4 | 5 | 3 | 4 | 2 |

**Conjunto 2.**

A continuación se presentan los pesos iniciales (X) y aumentos de peso (Y) de 10 ratas hembras de 28 a 84 días de edad, las cuales fueron sometidas a una dieta alta en proteínas.

| Número de rata | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Peso inicial en gramos | 50 | 64 | 76 | 64 | 74 | 60 | 69 | 68 | 56 | 48 |
| Peso final en gramos | 128 | 159 | 158 | 119 | 133 | 112 | 96 | 126 | 132 | 118 |

**Conjunto 3.**

Suponga una encuesta en donde se revisaron los archivos de la policía en un país y se anotó por ciudad el número de policías activos y el número de robos que fueron reportados en los últimos 5 años, los datos colectados se presentan a continuación:

| Ciudad | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Número de policías (X) | 64 | 53 | 67 | 52 | 82 | 59 | 67 | 90 | 50 | 77 | 88 | 71 | 58 |
| Número de robos (Y) | 625 | 750 | 560 | 690 | 515 | 680 | 630 | 510 | 800 | 550 | 550 | 525 | 625 |

**Conjunto 4.**

El gerente de una cadena de supermercados desea investigar la relación entre el número de empleados (X) y las ventas semanales en miles de quetzales (Y). Para ello toma una muestra de 15 tiendas con características semejantes, obteniendo las siguientes observaciones:

| Supermercado | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X | 10 | 17 | 17 | 17 | 19 | 24 | 24 | 30 | 32 | 32 | 36 | 40 | 41 | 43 | 48 |
| Y | 4 | 6 | 6 | 5 | 9 | 8 | 12 | 12 | 11 | 13 | 15 | 16 | 16 | 17 | 20 |

2.2 Para cada uno de los ejemplos que presentó en el Ejercicio 1 inciso 1, indique con palabras que relación esperaría (directa o inversa), es decir, a valores altos de la variable X que respuesta se esperaría en Y.

## 8. ESTIMACIÓN DE LOS PARÁMETROS DEL MODELO

Una vez que la evidencia exploratoria visual de los datos sugieren que efectivamente estos pueden representarse por una línea recta, se realiza el ajuste, es decir, se obtiene la ecuación de la recta de la regresión estimada.

$$\hat{y} = a + bx; \quad \text{en donde:}$$

$\hat{y}$ : valor estimado por la ecuación.

a: punto en donde la recta corta al eje vertical.

b: coeficiente de regresión o pendiente de la recta.

x: valores de la variable independiente o regresora.

Antes de entrar al procedimiento de estimar los valores de **a** y **b**, se presentarán dos métodos que intentan hacer esta estimación. Un método de estimación que a veces se utiliza en forma preliminar consiste en utilizar el diagrama de puntos o dispersión para dibujar la recta que mejor parezca representar la tendencia de los datos, o sea, que es un método gráfico. Este método puede presentar resultados aceptables, sobre

todo si la tendencia de los datos es muy marcada, pero tiene los inconvenientes de ser subjetivo, impreciso y sin ningún valor desde el punto de vista de la inferencia estadística.

El método analítico mas usado es el llamado **Método de Mínimos Cuadrados**, el cual tiene algunos años de antigüedad habiendo sido usado por Carl Gauss (1777-1855). La idea fundamental de este método es producir estimadores **a** y **b** que minimicen la suma de cuadrados de las distancias entre los valores observados (y) y los valores estimados ($\hat{y}$), esto es, que minimicen la suma de cuadrados de las longitudes de los segmentos de las líneas verticales que unen los datos observados con la recta estimada en la gráfica de dispersión (ver Figura 4).

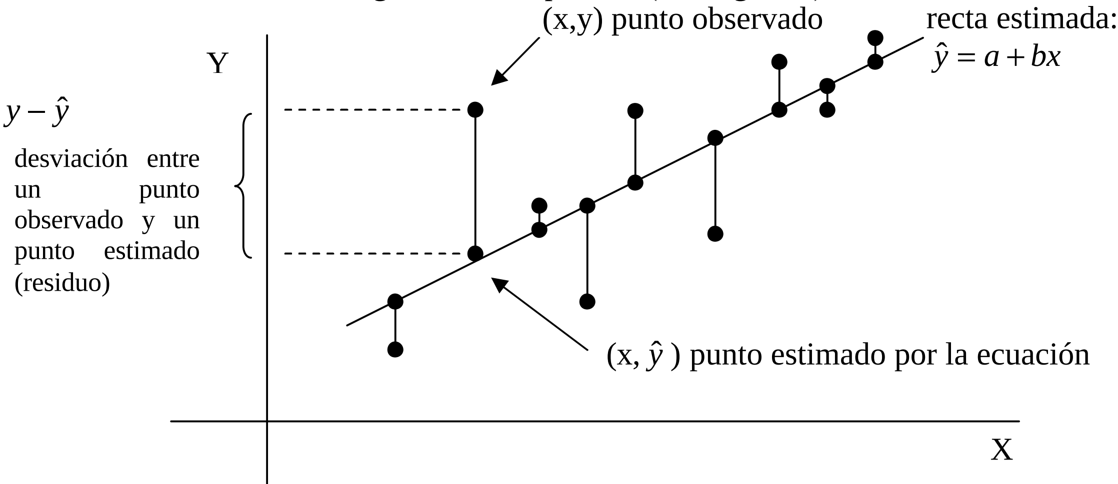



**Figura 4. Desviaciones entre los puntos observados y la recta de regresión estimada**.

Los estimadores resultantes de este método son ampliamente usados y tienen el respaldo teórico matemático-estadístico que les permiten tener buenas propiedades, desde el punto de vista de la construcción de estimadores.

Las ecuaciones para los estimadores **a** y **b** se obtienen mediante técnicas de cálculo diferencial que no se discutirán por el momento. Los resultados se presentan en seguida:

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \quad ; \qquad \text{(ecuación 1)}$$

$$a = \bar{y} - b\bar{x} \quad ; \qquad \text{(ecuación 2)}$$

en donde:

b: coeficiente de regresión.

$\sum xy$: suma de cuadrados cruzados de X e Y.

$\sum x$: suma de las observaciones de X.

$\sum y$: suma de las observaciones de Y.

$\sum x^2$: suma de los cuadrados de los valores de X.

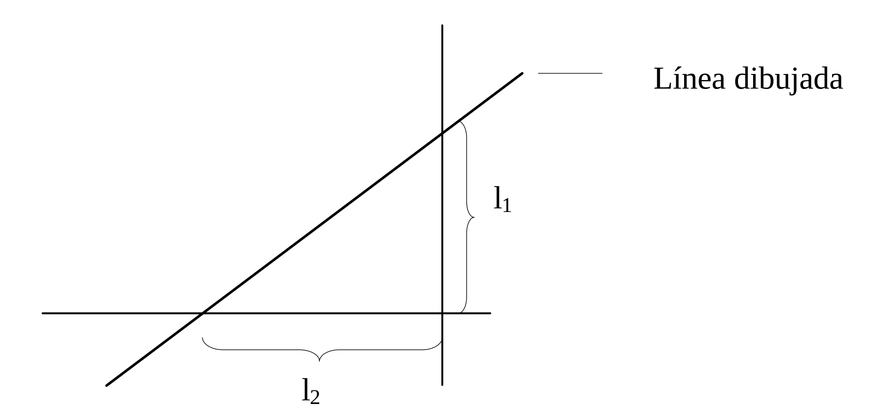
n: número de parejas que intervienen en los cálculos.

a: intercepto al origen.

$\bar{y}$: el promedio de las observaciones de la variable dependiente.

$\bar{x}$: el promedio de las observaciones de la variable independiente.

Ahora dibuje una línea recta sobre el diagrama de puntos que usted elaboró (Figura 3). Intente hacer una estimación geométrica de **a** y **b**. Recuerde que **a** es el punto en donde la recta corta al eje vertical, para este caso prolongue la línea dibujada hasta que corte al eje y anote el valor del punto. Y, **b** es la pendiente de la recta, para su cálculo trace un triángulo recto en donde la hipotenusa será el segmento de la recta dibujada y los lados serán las líneas paralelas a los ejes X e Y. Así:




Mida el largo de las líneas $l_1$ y $l_2$ (según la escala empleada), entonces $b = \frac{l_1}{l_2}$, este será el coeficiente de regresión. Escriba a continuación la ecuación que encontró:

$$precio = \frac{\qquad\qquad}{valor\ de\ a} + \frac{\qquad\qquad}{valor\ de\ b}\ \text{tamaño}$$

Usando la ecuación 1 y 2 ahora se encontrarán los estimadores mínimo cuadráticos, para lo cual se debe completar la siguiente tabla, con los cálculos sugeridos.

**Tabla 2. Valores para el cálculo de los estimadores**

| **No. Casa** | **Tamaño (x 100 m$^2$) X** | **Precio en miles de Q. Y** | **XY** | **X²** | **Y²** |
|---|---|---|---|---|---|
| **1** | 1,8 | 32 | 1.8*32=57.6 | 1.8*1.8=3.24 | 32*32=1024 |
| **2** | 1 | 24 | | 1 | 576 |
| **3** | 1,7 | 27 | 45,9 | | 729 |
| **4** | 2,8 | 47 | 131,6 | 7,84 | 2209 |
| **5** | 2,2 | 35 | 77 | 4,84 | |
| **6** | 0,8 | 17 | 13,6 | 0,64 | 289 |
| **7** | 3,6 | 52 | | 12,96 | 2704 |
| **8** | 1,1 | 20 | 22 | | 400 |
| **9** | 2 | 38 | | 4 | 1444 |
| **10** | 2,6 | 45 | 117 | 6,76 | |
| **11** | 2,3 | 44 | 101,2 | 5,29 | 1936 |
| **12** | 0,9 | 19 | 17,1 | | 361 |
| **13** | 1,2 | 25 | | 1,44 | 625 |
| **14** | 3,4 | 50 | 170 | 11,56 | |
| **15** | 1,7 | 30 | 51 | 2,89 | 900 |
| **16** | 2,5 | 43 | 107,5 | 6,25 | 1849 |
| **17** | 1,4 | 27 | 37,8 | | 729 |
| **18** | 3,3 | 50 | | 10,89 | 2500 |
| **19** | 2,2 | 37 | 81,4 | | 1369 |
| **20** | 1,5 | 28 | 42 | 2,25 | 784 |
| **Totales** | **40** | **690** | **1554.9** | **93.56** | **26178** |
| | **? X** | **? Y** | **? XY** | **? X$^2$** | **? Y$^2$** |

Otros cálculos:

n = 20; $\overline{X} = \frac{40}{20} = 2$ y $\overline{y} = \frac{690}{20} = 34.5$

Por lo tanto:

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \qquad \text{(ecuación 1)}$$

Al sustituir valores se tiene que:

$$b = \frac{1554.9 - \frac{(40)(690)}{20}}{93.56 - \frac{(40)^2}{20}} = 12.8982 \text{ y,}$$

$$a = \overline{y} - b\overline{x} \qquad \text{(ecuación 2)}$$

Sustituyendo con valores:

$$a = 34.5 - (12.8982)(2) = 8.7035$$

Uniendo los resultados, finalmente se encuentran los valores de la ecuación mínimo cuadrática de regresión lineal estimada:

| precio = 8.7035 + 12.8982(tamaño) |
|---|

Compare esta ecuación con la calculada geométricamente por usted. Comente:

________________________________________________

________________________________________________

________________________________________________

**EJERCICIO 3:**

3.1 Para cada uno de los cuatro conjuntos de datos presentados en el ejercicio 2 inciso 2.1 construya las ecuaciones de regresión empleando el método geométrico y los estimadores mínimo cuadráticos.

# 9. PRUEBA DE ADECUACION DEL MODELO Y ANÁLISIS DE RESIDUOS.

Una cantidad que indica que tan buena es la recta de regresión estimada, es el coeficiente de Determinación ($R^2$). El coeficiente de Determinación $R^2$ es un indicador de la variabilidad en Y que es explicada por el modelo. El coeficiente de Determinación varia entre 0 y 1 ($0 \le R^2 \le 1$); entre mas cerca este de uno es mejor el ajuste del modelo a los datos. No existe frontera para clasificar con base en $R^2$ los modelos en buenos y malos. Todo depende del problema en turno que se este resolviendo.

La expresión para calcular el valor de $R^2$ está definida conceptualmente como:

$$R^2 = \frac{\text{variación explicada por la Regresión}}{\text{variación total de las observaciones de } Y}; \quad \text{y algebraicamente:}$$

$$R^2 = \frac{SSR}{SST}; \qquad \text{(ecuación 3)}$$

en donde:

$$SSR = \frac{\left[\sum xy - \frac{(\sum x \sum y)}{n}\right]^2}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$SST = \sum y^2 - \frac{(\sum y)^2}{n}$$

SSR : suma de cuadrados de la Regresión.

SST : suma de cuadrados del total.

Es oportuno indicar que el $R^2$ será confiable o creíble siempre que existan por lo menos 10 datos (parejas) por cada parámetro que se desea estimar en el modelo. Por tal razón a continuación se escribe la expresión para calcular el coeficiente de Determinación ajustado o corregido ($R^2_A$):

$$R_A^2 = 1 - \frac{n-1}{n-p}(1 - R^2) \qquad \text{(ecuación 4) ;}$$

en donde:

$R^2_A$: coeficiente de Determinación ajustado.

n : número de parejas.

p: número de parámetros estimados en el modelo.

$R^2$: coeficiente de Determinación.

Ahora, continuando con el ejemplo que a servido en el seguimiento del tema, se tienen los resultados, empleando las ecuaciones 3 y 4:

$$SSR = \frac{\left[1554.9 - \frac{(40)(690)}{20}\right]^2}{93.56 - \frac{(40)^2}{20}} = \frac{(174.9)^2}{13.56} = 2255.9004$$

$$SST = 26178 - \frac{690}{20}^2 = 2373$$

por lo tanto, $R^2 = \frac{2255.9004}{2373} = 0.9506$, antes de interpretar el $R^2$, se calculará el $R^2_A$ para tener un mejor criterio.

$$R_A^2 = 1 - \frac{20-1}{20-2}(1-0.9506) = 1 - \frac{19}{18}(0.0493) \approx 0.9480 .$$

Se observó que el valor cambio levemente, por lo tanto, se puede decir que el modelo explica 0.9480 (94.80%) de la variabilidad contenida en los datos, es candidato a ser buen modelo.

Sin embargo, es necesario hacer un estudio mas detallado de la variabilidad, por lo que debe construirse una tabla de análisis de varianza.

**EJERCICIO 4:**

4.1 Calcule para cada uno de los conjuntos de datos del ejercicio 2 inciso 2.1, los correspondientes valores de $R^2$ y $R^2_A$; comente los resultados.

# 10. ANÁLISIS DE RESIDUOS

Previo a entrar de lleno al análisis de varianza se considera oportuno efectuar un análisis de los residuos del modelo. La técnica de análisis de residuos es útil para revisar aspectos importantes de la modelación con regresión, entre otros:

- Revisión del cumplimiento de las suposiciones que hacen válido el análisis de varianza: Homogeneidad de varianza, independencia de errores y distribución normal de los residuos.
- Sugerencia de otros modelos de regresión.
- Determinación de valores atípicos.

Las propiedades teóricas de los residuos permiten que su utilización sea práctica, es decir, que solamente mediante una gráfica se pueden obtener resultados concluyentes. Se define como residuo a la diferencia entre el valor observado de Y y el valor predicho por ajuste, es decir, $\hat{e}_i = y_i - \hat{y}_i \quad ; \quad i = 1,...,n$, donde $y_i$ es el valor observado, $\hat{y}_i$, valor estimado por la ecuación.

Para el problema que se a tomado como referencia se tiene:

La casa 1 se tiene un tamaño ($x_1$) de 1.8, precio de venta ($y_1$) de 32; para el cálculo del valor estimado habrá que sustituir el valor del tamaño (1.8) en la ecuación mínimo cuadrático ajustada, precio = 8.7035 + 12.8982 (1.8) = 31.92. Por lo tanto, el residuo correspondiente será $\hat{e}_1 = 32 - 31.92 = 0.08$. Igualmente para la casa 2 y 3 los residuos serán $\hat{e}_2 = 2.40\ y\ \hat{e}_3 = -3.63$, respectivamente. A continuación complete la Tabla 3 que contiene los datos observados, residuos y los estimados del problema relacionado al valor de las casas.

**TABLA 3. Residuos de los precios de las veinte casas.**

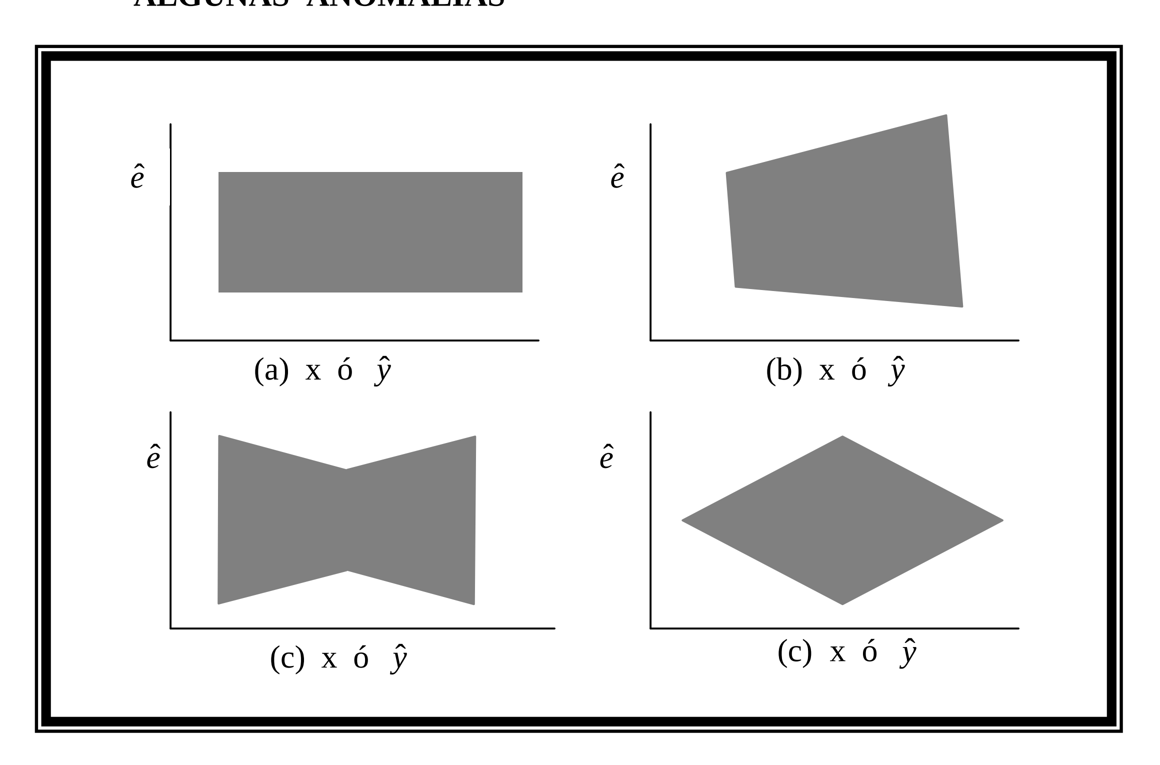
| No. Casa | Tamaño (x 100 m$^2$) X | Precio en miles de Q. $y_i$ | Precio Estimado $\hat{y}_i$ | Residuos $(y_i - \hat{y}_i)$ |
|---|---|---|---|---|
| 1 | 1,8 | 32 | 31,92 | 0,08 |
| 2 | 1 | 24 | 21,60 | 2,40 |
| 3 | 1,7 | 27 | 30,63 | -3,63 |
| 4 | 2,8 | 47 | 44,82 | |
| 5 | 2,2 | 35 | 37,08 | -2,08 |
| 6 | 0,8 | 17 | 19,02 | -2,02 |
| 7 | 3,6 | 52 | 55,14 | -3,14 |
| 8 | 1,1 | 20 | 22,89 | |
| 9 | 2 | 38 | 34,50 | 3,50 |
| 10 | 2,6 | 45 | 42,24 | 2,76 |
| 11 | 2,3 | 44 | 38,37 | |
| 12 | 0,9 | 19 | 20,31 | -1,31 |
| 13 | 1,2 | 25 | 24,18 | 0,82 |
| 14 | 3,4 | 50 | 52,56 | -2,56 |
| 15 | 1,7 | 30 | 30,63 | -0,63 |
| 16 | 2,5 | 43 | 40,95 | |
| 17 | 1,4 | 27 | 26,76 | 0,24 |
| 18 | 3,3 | 50 | 51,27 | -1,27 |
| 19 | 2,2 | 37 | 37,08 | -0,08 |
| 20 | 1,5 | 28 | 28,05 | -0,05 |

Con la tabla completa se pueden graficar los residuos, para lo cual habrá que colocar en el eje de las X's ya sea los valores predichos o los valores de la variable regresora X y en el eje vertical los residuos. Se han establecido algunos patrones de comportamiento que ayudan a identificar o mostrar algunos resultados que están relacionados con el cumplimiento de las suposiciones teóricas del análisis de

varianza e identificación de modelos. Gráficamente, dado que entre los valores predichos o los de la variable X y los residuos no debe existir relación alguna, cualquier patrón diferente de uno aleatorio será indicativo de alguna patología.

Para juzgar este tipo de gráficos se presentan en la Figura 5 una serie de patrones anómalos.

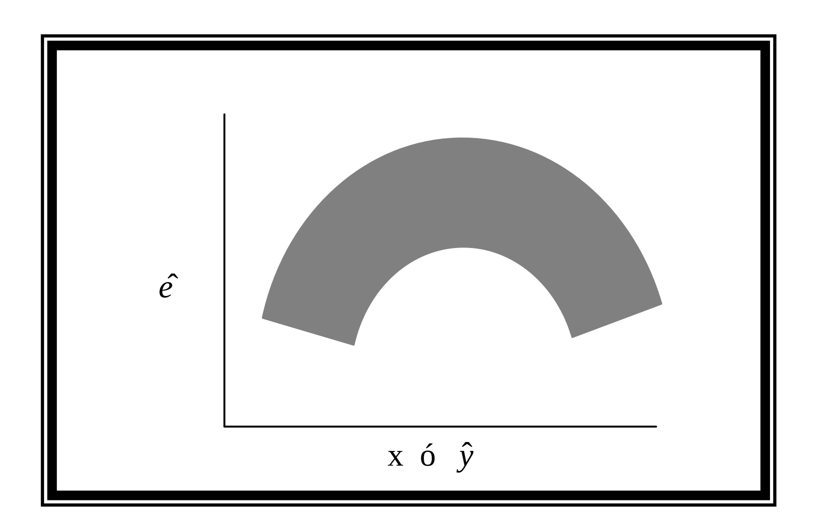
**FIGURA 5 PATRONES INDICATIVOS DE UN BUEN MODELO Y ALGUNAS ANOMALIAS**




La figura 5(a) muestra un **comportamiento aleatorio** representante de un buen modelo. Cuando la **suposición de homogeneidad de varianza (homocedasticidad)** no se cumple, el patrón esperado será el mostrado en las figuras 5(b), 5(c), ó 5(d).

Otro ejemplo importante que se puede explorar a través de las gráficas de residuos es el efecto **adicional (curvilíneo) de la X** que no ha sido incorporado al modelo. Por ejemplo si la X tiene un efecto cuadrático, al graficar los residuos contra X o los valores estimados, éste aparecerá como se muestra en la Figura 6.

**FIGURA 6. Gráfica de residuos para un modelo cuadrático**

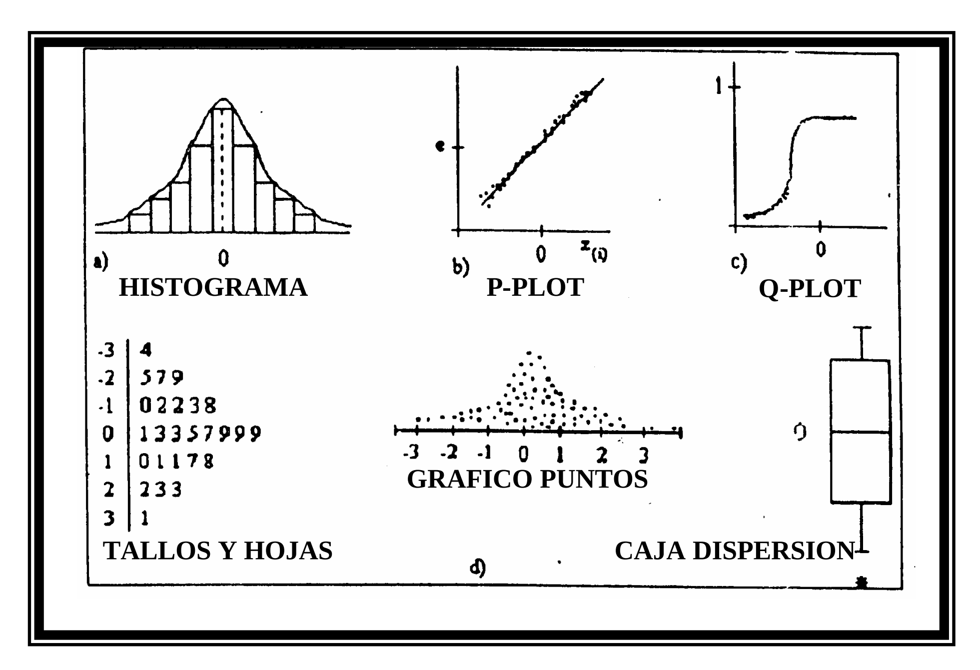



Si el modelo lineal fue el propuesto originalmente, agregando un término cuadrático se corrige este patrón. Una suposición que se hace en el análisis de varianza es la **independencia de los errores** la cual para ser verificada se necesita conocer exactamente como fueron obtenidos los datos, o sea, conocer la definición de la muestra y principalmente como fue su planeamiento.

Otro aspecto importante relativo a los supuestos es la **normalidad**. La normalidad se requiere para garantizar la eficiencia de las pruebas de hipótesis y aunque no es un supuesto muy importante, ya que un tamaño de muestra grande puede atenuar los problemas que surgen por no cumplirse esta suposición. Para tener una idea de la razonabilidad de este supuesto podemos explorar los residuos a través de gráficos como histogramas con curva ajustada, p-plot, q-plot, diagramas de tallos y hojas, diagramas de dispersión, entre otros.

A continuación se presentan en la Figura 7 cada uno de estos gráficos y diagramas, cuando los datos tienen una apariencia de normalidad razonable.

**FIGURA 7. GRAFICAS Y DIAGRAMAS CON APARIENCIA DE NORMALIDAD RAZONABLE**




**OBSERVACIONES ATÍPICAS.**

Otro problema que puede afectar a la bondad de ajuste del modelo es la presencia de observaciones atípicas. A veces la atipicidad de un dato se observa en un gráfico de dispersión, otras veces, es necesario ajustar el modelo y observar los residuos para identificarlo. Existen varios criterios basados en varios tipos de residuos, que pueden guiar en la identificación concreta de los puntos atípicos, estos tipos de residuos son:

a. Residuos crudos ($\hat{e}_i = y_i - \hat{y}_i$).

b. Residuos estandarizados ($d_i = \frac{\hat{e}_i}{\sqrt{CME}}$) que siguen una distribución aproximadamente normal (CME, es el cuadrado medio del error).

c. Residuos estudentizados ($r_i = \frac{\hat{e}_i}{SE_{(-i)}\sqrt{1-h_{ii}}}$) donde $h_{ii}$ es el valor de la matriz Hat.

Para "medir" la influencia se han propuesto varios criterios para declarar una observación atípica:

1. Sí $h_{ii} > \frac{2(k+1)}{n}$; donde k es el número de variables regresoras.

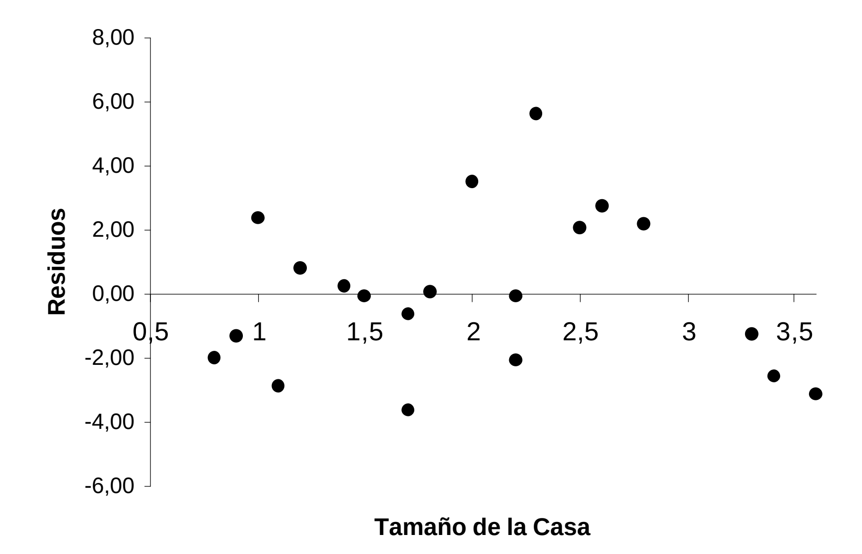
2. Sí $|d_i| > 3$.

3. Sí $D_i = \frac{r_i}{k+1} * \frac{h_{ii}}{1-h_{ii}} > 1$.

Para el caso de regresión múltiple es preferible usar el tercer criterio, llamado de la D de Cook.

Para el caso del ejercicio de referencia se tiene el siguiente gráfico de residuos el cual se observa en la figura 8.

**Figura 8. Gráfica de los valores de tamaño de la casa y los residuos.**




Al observar el gráfico de residuales, que se puede concluir con respecto a:

1. Los residuos graficados insinúan un candidato a buen modelo? ¿Por qué?

____________________________________________________________

____________________________________________________________

2. Cuál es su opinión con respecto a la varianza de los errores (homogeneidad)?

____________________________________________________________

____________________________________________________________

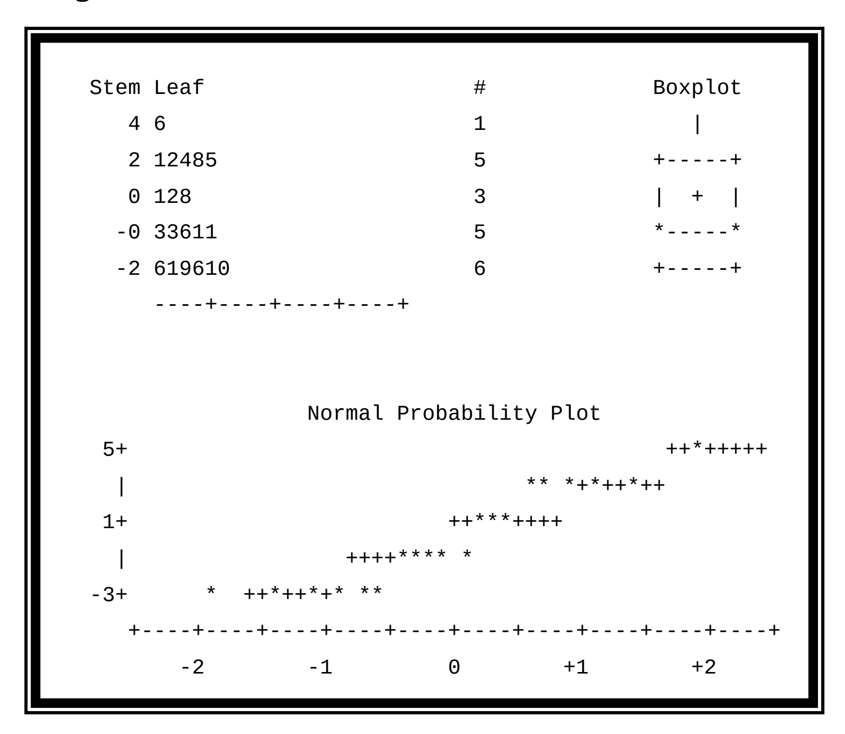
3. Es necesario agregar un término cuadrático al modelo? ¿Por qué?

____________________________________________________________

____________________________________________________________

En la figura 8 se presenta el diagrama de tallos y hojas y el diagrama de puntos de los residuos del problema.

**Figura 8 Diagrama de la distribución de los Residuos**



```
Stem Leaf                     #             Boxplot
   4 6                        1                |
   2 12485                    5             +-----+
   0 128                      3             |  +  |
  -0 33611                    5             *-----*
  -2 619610                   6             +-----+
     ----+----+----+----+


                Normal Probability Plot
 5+                                          ++*+++++
  |                                ** *+*++*++
 1+                          ++***++++
  |                  ++++**** *
-3+       *  ++*++*+* **
   +----+----+----+----+----+----+----+----+----+----+
       -2        -1         0        +1        +2
```

Cuando el análisis gráfico de los residuales no muestra claramente la normalidad, existen pruebas numéricas para este mismo propósito, una de ellas, la prueba de Shapiro-Wilks concluye que los residuos en análisis si tienen una distribución normal (W:Normal = 0.961673, Pr < W = 0.5807).

Cree razonable la distribución normal de los residuos? ¿Por qué?

____________________________________________________________

____________________________________________________________

El análisis de la presencia o no de observaciones atípicas se deja a criterio del lector lo cual puede realizar después de haber concluido el análisis de varianza.

**EJERCICIO 5:**

5.1 Para cada uno de los conjuntos de datos propuestos en el ejercicio 2, inciso 2.1, se le pide efectuar el correspondiente análisis de residuos y escribir sus conclusiones.

## 11. ANÁLISIS DE VARIANZA (ANOVA).

En la evaluación de la adecuación del modelo y análisis de residuos, específicamente cuando se presentó en Coeficiente de Determinación ($R^2$), se indicó que este no era suficiente para juzgar al modelo, dado que había que hacer un estudio mas detallado de la variabilidad, entonces, se hará ese estudio a continuación.

Cuando las suposiciones de independencia, normalidad y homogeneidad de las varianzas de los residuos han sido verificadas, puede entonces, iniciarse el análisis de varianza. El ANOVA de la regresión es una herramienta estadística para estudiar la relación entre la variable respuesta o dependiente y la variable predictora o explicatoria. Para realizar el ANOVA se propone el siguiente procedimiento:

**Paso 1.** **Planteamiento de la Hipótesis**:

En esta parte debe especificarse la hipótesis que se someterá a prueba, para este caso:

$$H_o: \quad b_1 = 0 \quad \text{(no hay regresión)}$$

$$H_a: \quad b_1 \neq 0 \quad \text{(si hay regresión)}$$

**Paso 2.** **Tabla de resultados**:

Los resultados necesarios para los cálculos aparecen en la Tabla 2 de este documento.

**Paso 3.** **Cálculos**:

- **Grados de libertad**:

Grados de libertad para la regresión: glr = 1

Grados de libertad para el error: gle = n – 2

Grados de libertad para el total: glt = n – 1

- **Cálculo de Sumas de Cuadrados**:

Suma de cuadrados para la regresión (SSR)

$$SSR = \frac{\left[\sum xy - \frac{\sum x \sum y}{n}\right]^2}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Suma de cuadrados para el total (SST)

$$SST = \sum y^2 - \frac{(\sum y)^2}{n}$$

Suma de cuadrados para el error (SSE)

SSE = SST – SSR

- **Cálculo de cuadrados medios**:

Cuadrado medio de la regresión (CMR)

$$CMR = \frac{SSR}{glr}$$

Cuadrado medio del error (CME)

$$CME = \frac{SSE}{gle}$$

- **Cálculo del estadístico F:**

$$F = \frac{CMR}{CME}$$

Los cálculos anteriores pueden resumirse como se muestra en la Tabla 4.

**Tabla 4. Tabla general del análisis de varianza para el modelo de regresión lineal simple.**

| **Fuente de Variación** | **Grados de Libertad** | **Suma de Cuadrados** | **Cuadrado Medio** | $\mathbf{F_c}$ |
|---|---|---|---|---|
| Regresión | 1 | SSR | $CMR = \frac{SSR}{1}$ | $F_c = \frac{CMR}{CME}$ |
| Error | n - 2 | SSE | $CME = \frac{SSE}{n-2}$ | |
| Total | n – 1 | SST | | |

Donde:

$F_c$: Estadístico calculado de F.

El valor de $F_c$ tiene asociado un p-value que indica la evidencia probabilística para rechazar la $H_o$. Los criterios específicos para las pruebas de hipótesis a partir de la observación del p-value son los siguientes:

a. Sí $p > 0.1$.Se declara que no existe evidencia suficiente para rechazar $H_o$.

b. Sí $0.05 < p \leq 0.1$. Se declara que hay evidencia, pero baja, es decir, se rechaza $H_o$ con baja significancia.

c. Sí $0.01 < p \leq 0.05$. Se dice que existe suficiente evidencia para rechazar $H_o$. Se rechaza $H_o$ con evidencia significativa.

d. Sí $p \leq 0.01$. Se dice que existe evidencia altamente significativa para rechazar $H_o$.

Se aprovecha el momento para indicar que esta prueba es equivalente a observar el intervalo de confianza al 95% o al 99% para $b_1$ la regla de decisión en tal caso consiste en observar si el intervalo contiene a cero, en cuyo caso no se rechaza la $H_o$ (Hipótesis Nula), con el nivel de significancia especificado en el intervalo, en caso contrario la $H_o$ se rechaza con la significancia correspondiente.

Para el ejemplo que se desarrolla paralelamente en este documento la Tabla de ANOVA se muestra a continuación, buena parte de las operaciones ya se realizaron cuando se calculó el coeficiente de determinación ($R^2$), por lo que ahora solo se muestran los resultados.

**Tabla 5. ANOVA para el tamaño y precio de las casas.**

| Fuente de Variación | Grados de Libertad | Suma de Cuadrados | Cuadrado Medio | $F_c$ | p - value |
|---|---|---|---|---|---|
| Regresión | 1 | 2255.900 | 2255.900 | 346.767 | 0.0001 |
| Error | 18 | 117.01 | 6.506 | | |
| Total | 19 | 2373.00 | | | |

Con estos resultados se le pide plantear la hipótesis sometida a prueba y escribir sus conclusiones:

______________________________________________________________________

______________________________________________________________________

______________________________________________________________________

**EJERCICIO 6.**

6.1 Para cada unos de los conjuntos de datos propuestos en el Ejercicio 2, inciso 2.1, se le pide realizar el correspondiente ANOVA y escribir sus conclusiones.

## 12. INTERPRETACIÓN Y USO DEL MODELO PARA EL PROPÓSITO ESTABLECIDO.

Cuando los supuestos del Anova se han verificado, el Anova ha sido significante, el análisis de residuos muestra un patrón deseado y el Coeficiente de Determinación es razonablemente alto, entonces, se puede decir que el modelo construido está en condiciones de usarse para el propósito establecido previamente.

Es oportuno tener presente algunas sugerencias para hacer un uso adecuado del modelo, entre otras:

1. Un buen modelo de regresión no implica necesariamente una relación causa-efecto entre las variables. Para establecer causalidad, la relación entre variables regresoras y las variables de respuesta deben tener un fundamente físico y lógico, es decir, la relación debe ser sugerida por consideraciones teóricas.
2. Generalmente los modelos de Regresión son válidos únicamente en la región o rango de valores que fueron observados en la variable independiente. Es decir, no se sugiere hacer extrapolaciones.
3. Los datos que se incluyen en el análisis de Regresión deben ser representativos del sistema estudiado. Para que el modelo cumpla el propósito para el cual fue generado, la recolecta de datos debe planificarse de manera adecuada para que las conclusiones obtenidas de éste sean correctas.
4. Las predicciones hacerlas por medio de un intervalo de confianza.
5. Las conclusiones acerca del modelo y uso del mismo son válidas para sistemas similares al que pertenecen los datos que permitieron obtener el modelo.

Expuesto lo anterior, se procederá a interpretar el modelo. En términos generales el modelo estimado tendrá la siguiente forma:

$$\hat{y}_i = a + bx_i$$

Donde:

$\hat{y}_i$ : Valor estimado por la ecuación.

$a$ : Intercepto al origen de la recta estimada

$b$ : Coeficiente de Regresión

$x_i$ : Valor de la variable independiente.

## INTERPRETACIÓN DE LOS COEFICIENTES

Del modelo estimado habrá que interpretar “a” y “b”. Por lo tanto, el intercepto (a) según el contexto del problema puede tener dos connotaciones, la primera, **será el valor de la variable respuesta cuando la variable independiente tiene un valor de cero.** Y, la segunda, si el valor de cero de la variable independiente no tiene sentido lógico, entonces “a” **será un valor de ajuste sin sentido práctico.**

Ahora, "b" llamado coeficiente de Regresión o pendiente de la recta, tendrá la siguiente interpretación:

1. Si el valor de "b" es negativo: **por cada unidad de X que aumente, la variable dependiente disminuirá "b" unidades, o viceversa (regresión inversa).**
2. Si el valor de "b" es positivo: **por cada unidad de X que aumente, la variable dependiente también aumentará "b" unidades, o viceversa (regresión directa).**

Por lo tanto, si el modelo de regresión ajustado para el Ejercicio paralelo es:
Precio = 8.7935 + 12.8982(tamaño de la casa), la interpretación del mismo será:
Intercepto (8.7935):

________________________________________________________________

________________________________________________________________

Coeficiente de Regresión (12.8982):

________________________________________________________________

________________________________________________________________

Si se desea hacer una predicción del precio, cuando el tamaño de la casa es de 3 (x 1000 Quetzales), al usar el modelo (precio = 8.7935 + 12.8982(3)), el precio será de 47.3981 (x 1000 Quetzales).

La predicción por intervalo para este caso es:

$$\hat{y}_0 \pm t_{\frac{a}{2}} \sqrt{CME} \sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SSX}} \qquad \text{(ecuación 5)}$$

En donde:

$\hat{y}_0$ : Valor predicho

$t_{\frac{a}{2}}$ : Valor de t a cierto % de confianza

CME: Cuadrado medio del error (Ver Tabla 4)

n: Número de parejas

$x_0$ : Valor particular para la predicción

$\bar{x}$ : Promedio de los valores de la variable independiente

SSX: Suma de cuadrados corregidos de los valores de X $\left( SSX = \sum x^2 - \frac{\left(\sum x\right)^2}{n} \right)$

Sustituyendo valores en la Ecuación 5, se tiene:

$\hat{y}_0 = 47.3981$, $t_{\frac{a}{2}} = 2.10$ ( al 95% de confianza), CME = 6.506, n = 20, $x_0 = 3$,

$\bar{x} = 2$, y SSX = 13.56. $\left( SSX = 93.56 - \frac{40^2}{20} \right)$ (valores tomados de la Tabla 2).

Por lo tanto,

$47.3981 \pm 2.10\sqrt{6.506}\sqrt{1 + \frac{1}{20} + \frac{(3-2)}{13.56}}$ = $47.3981 \pm 2.10(2.55)(1.06) =$

$47.3981 \pm 5.67$ = **(41.73, 53.07),** cuya interpretación será:

Se espera con 95 % de confianza que el valor de una casa que mide 300 m$^2$ este comprendido entre Q 41,730 y Q 53,070.

**EJERCICIO 7**

7.1 De los cuatro conjunto de datos dados en el Ejercicio 2, inciso 2.1, para los modelos significativos, haga una predicción puntual por intervalo interpretando los resultados en el contexto del problema.